1. A Simulation on the Importance of Distributional Information

Using simulated response time data, we compare the following two "behavioral models" for estimating reliability: (1) the traditional two-stage summary approach (compute means, take the difference, and compute a test-retest correlation), and (2) a method that contrasts the distributions holistically (as articulated below) before computing a test-retest correlation. To generate simulated data, we drew response times from a lognormal distribution (right-skewed as in most response time tasks; Figure 2 of the main text) for each participant and condition, then compared test-retest correlations across the approaches. We simulated 150 "participants", each of whom completed the response time task at two different sessions, with an artificial "congruent" and "incongruent" condition at each timepoint. Critically, the parameters used to generate response time data at each timepoint were *exactly the same for each participant*—the generative parameters had true test-retest correlations of $r = 1.0$. The procedure produced right-skewed response time distributions with 80% of draws between 300 and 2000 milliseconds. The specific steps used to simulate data were as follows:

1. Draw the mean parameter of each simulated participant's lognormal distribution for the "incongruent" task condition from a group-level normal distribution N(-1.2, 0.2),
2. Add -.5 to each simulated participant's values from step 1 to calculate the mean parameter of the lognormal distribution for the "congruent" condition (i.e. the condition manipulation has the same underlying effect on all participants),
3. Draw the standard deviation parameter for each participant's lognormal distribution from a uniform distribution U(0,1), which is shared across both task conditions, and

4. Sample from each participant's lognormal distribution for each condition and timepoint using parameters from steps 1-3.

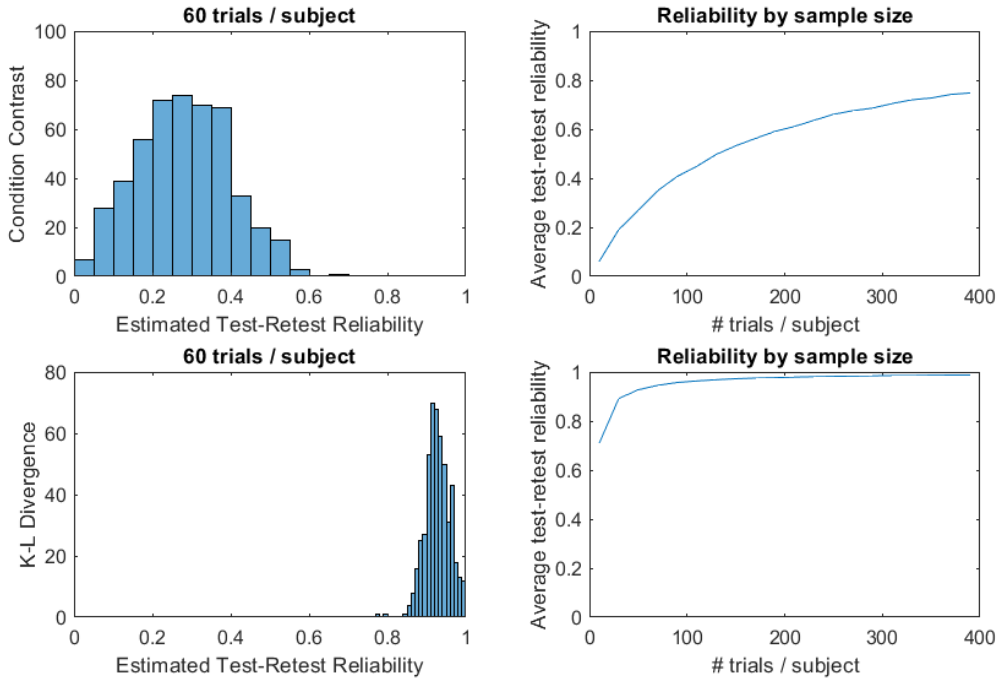
For step 4, we used various different sample sizes (i.e. number of trials within each condition) to determine how test-retest estimates were affected. Across sample sizes, steps 1-4 yielded typical right-skewed response time distributions with 80% of draws between 300 and 2000 milliseconds.

For each simulated participant, we conducted two reliability tests. The first simulated a traditional reliability analysis of performance (e.g., test-retest reliability of mean response time difference between congruent and incongruent trials in the Stroop task), with knowledge that the true generating parameters were unchanged across test and retest. For each of the two sessions, we computed the mean difference between each participant's "incongruent" and "congruent" response time distributions. Next, we estimated Pearson correlations between the Session 1 and Session 2 mean differences across participants as an index of test-retest reliability. We repeated this procedure 1,000 times at sample sizes ranging from 10 to 400 per participant.

Figure S1 shows results of this analysis. The top left panel shows an example distribution of inferred test-retest estimates across 1,000 repetitions for a sample size of 60 trials. These test-retest reliabilities of mean contrasts ranged from close to $r = 0$ to $r = .5$ (middle-left panel, Figure 4). Test-retest reliability improved substantially with more trials for each participant, to around $r = .8$ at 400 trials (middle-right panel).

Figure S1. Test-retest reliability simulations.

The panels compare the mean difference between two conditions (top), and contrasting distributions using K-L divergence (bottom). The left panels show estimated reliabilities for sample sizes of 60 response times per participant (a typical size for the IAT) across 1,000 simulations. The right panels show how average reliability of these contrasts changes across sample sizes.



To demonstrate how important distributional information can be, we performed a second reliability analysis which used Kullback-Leibler (K-L) divergence to quantify the relative difference between each participant's response time distributions across trials within conditions. K-L divergence is an information theoretic measure that quantifies the relative degree of difference between two continuous probability distributions, which allows us to compare two probability distributions holistically without making specific parametric assumptions regarding the shape of each distribution:

$$D_{KL}(P_1||P_2) = \int_{-\infty}^{\infty} p_1(x)\ \log\left(\frac{p_1(x)}{p_2(x)}\right) dx \qquad \text{(S1)}$$

Here, P_1 and P_2 are the full response time distributions for a given participant in conditions 1 (incongruent) and 2 (congruent), and $p_1(x)$ and $p_2(x)$ indicate the probability density of the response time distributions for conditions 1 and 2 at time x. Importantly, if $p_1(x) = p_2(x)$, the log term returns 0 for the given time x, which indicates that the probability densities are equivalent at time x. If $p_1(x) = p_2(x)$ for all values of x, then $D_{KL}(P_1||P_2) = 0$, indicating that the probability distributions over the response times in conditions 1 and 2 are identical. To the degree that $p_1(x) \neq p_2(x)$ across all values for x, $D_{KL}(P_1||P_2)$ becomes increasingly positive.

We estimated $D_{KL}(P_1||P_2)$ for each participant by first passing a (Gaussian) kernel density estimator over each of their response times for the congruent and incongruent conditions and first and second time points, giving estimated probability densities for four distributions (time 1 / congruent, time 1 / incongruent, time 2 / congruent, and time 2 / incongruent). We then estimated test-retest as the Pearson's correlation of the $D_{KL}(P_1||P_2)$ measure, as opposed to a mean contrast, across the simulated timepoints for each of the 1,000 repetitions, the results of which are shown in Figure S1.

Results show that most test-retest reliabilities based on K-L divergence between congruent and incongruent trials were between $r = .85$ and 1.0. Use of a distribution-informed metric was therefore much more successful in recovering the true test-retest of reliability ($r = 1.0$), which has both empirical and theoretical implications for analyzing behavioral data. Empirically, achieving desirable psychometric properties such as high test-retest reliabilities requires many behavioral observations (trials) from each participant—particularly when relying on traditional behavioral models (e.g., mean contrasts). Indeed, the reliability of the mean contrasts only began to approach $r = .8$ after 400 trials per participant per condition, which is far beyond the typical number of trials used in such tasks. Theoretically, the implications are much broader. Psychometric properties of behavioral paradigms are highly dependent on underlying behavioral models (e.g., mean contrasts versus K-L divergence). Accordingly, models that are sensitive to the entire distribution of individual-level behavior are better suited for recovering individual differences. For response times, this necessitates behavioral models that capture full distributions of response times across trials, and the right-skewed nature often observed for such distributions (e.g., Heathcote et al., 1991; Hockley & Corballis, 1982; Kvam & Busemeyer, 2020; Leth-Steensen et al., 2000; Whelan, 2008). For dichotomous or categorical data, as we will demonstrate with the delay discounting task, this requires models that produce probabilities that represent how likely participants are to select each of the possible responses.

2. Maximum Likelihood Estimation as the Two-stage Approach

As noted in the main text, the two-stage approach of estimating mean response times, contrasting them, and then entering the results into a secondary model can be viewed through the lens of maximum likelihood estimation. Specifically, assuming that response times arise from a

normal generative model, the sample mean and standard deviation are the analytic maximum likelihood estimators for the normal mean and standard deviation generative parameters. Therefore, the mean contrast approach is equivalent to contrasting the maximum likelihood estimates within participants across conditions in the response time tasks. By entering the maximum likelihood point estimates into a secondary statistical model, the uncertainty associated with each estimate (which can be gleaned through the Fisher information matrix) is ignored—akin to the two-stage summary statistic approach. This correspondence between the summary statistic approach and the use of maximum likelihood estimates motivates our use of maximum likelihood for the delay discounting model (to compare to the generative modeling approach), which we expand on below.

The hyperbolic model that we used to fit the delay discounting task data in the main text (more details below in section 3) is a generative model at the level of individual behavior (i.e. it can generate distributions of trial-level choices consistent with observed behavior). However, discounting rates from the hyperbolic model are traditionally estimated individually for each participant by fitting a hyperbolic curve (i.e. Equation S9 described in section 3) to their indifference points using optimization methods such as least squares estimation, after which the point estimates are used in subsequent analyses (see Odum, 2011, p. 431). This approach ignores uncertainty with respect to both the indifference points as well as the probabilistic nature of trial-by-trial responses (captured by Equation S10 described in section 3). Similarly, some studies fit the full model (both Equations S9 and S10 described in section 3) using maximum likelihood estimation and conduct subsequent analyses using the resulting point estimates, which is akin to the two-stage sample mean/standard deviation contrast approach often used to analyze response time tasks. Regardless of the procedure, reducing the discounting rate and choice sensitivity

parameters to single point estimates ignores important sources of individual-level uncertainty, which means that a secondary statistical model relying on such estimates would not be considered a full generative model of the effect of interest (e.g., group difference, correlation, test-retest, etc.). Therefore, we compare the two-stage maximum likelihood approach to the generative modeling approach for the delay discounting task, the results of which are presented in Figure 7 in the main text, and in Figure S6 in section 6 below.

3. Full Specification of Generative Models

3.1 Response Time Models

To facilitate test-retest analysis for the response time modes, the means and standard deviations for the normal, lognormal and shifted lognormal models (Equations 4, 5, and 6 in the main text, respectively) can be re-parameterized such that each participant is characterized by a "baseline" mean and standard deviation in condition 1 (e.g., the congruent condition, or $c = 1$), with an added "change" parameter that reflects the differences for condition 2 (e.g., the incongruent condition, or $c = 2$):

$$\mu_{i,c,t} = \begin{cases} \mu_{i,\text{base},t}, & \text{if } c = 1 \\ \mu_{i,\text{base},t} + \mu_{i,\Delta,t}, & \text{if } c = 2 \end{cases}$$
$$\sigma_{i,c,t} = \begin{cases} \exp(\sigma_{i,\text{base},t}), & \text{if } c = 1 \\ \exp(\sigma_{i,\text{base},t} + \sigma_{i,\Delta,t}), & \text{if } c = 2 \end{cases} \quad \text{(S2)}$$

Here, $\mu_{i,base,t}$ and $\sigma_{i,base,t}$ simply reflect the mean and standard deviation for each participant i in the congruent condition at time t, while $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$ represent changes in the corresponding means and standard deviations between conditions—in other words, participants' "Stroop effects". Note that the standard deviation parameters are exponentially transformed so that they are non-negative.

To estimate test-retest reliability, we can assume that individual-level change scores (i.e. $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$) are correlated. As described in the main text, we can estimate their correlation by assuming they are drawn from multivariate normal distributions as opposed to independent normal distributions (Equation 7 of the main text):

$$\begin{bmatrix}\mu_{i,\Delta,1}\\ \mu_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\mu_{\text{mean},\Delta,1}\\ \mu_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_\mu\right)$$
$$\begin{bmatrix}\sigma_{i,\Delta,1}\\ \sigma_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\sigma_{\text{mean},\Delta,1}\\ \sigma_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_\sigma\right) \quad \text{(S3)}$$

The use of a multivariate normal distribution allows us to estimate covariances (the $\mathbf{S}_\mu$ and $\mathbf{S}_\sigma$ matrices) between the individual-level parameters across timepoints, which can be decomposed into the group-level parameter variances and their correlations:

$$\mathbf{S}_\mu = \begin{pmatrix}\mu_{\text{sd},\Delta,1} & 0\\ 0 & \mu_{\text{sd},\Delta,2}\end{pmatrix} \mathbf{R}_\mu \begin{pmatrix}\mu_{\text{sd},\Delta,1} & 0\\ 0 & \mu_{\text{sd},\Delta,2}\end{pmatrix}$$
$$\mathbf{S}_\sigma = \begin{pmatrix}\sigma_{\text{sd},\Delta,1} & 0\\ 0 & \sigma_{\text{sd},\Delta,2}\end{pmatrix} \mathbf{R}_\sigma \begin{pmatrix}\sigma_{\text{sd},\Delta,1} & 0\\ 0 & \sigma_{\text{sd},\Delta,2}\end{pmatrix} \quad \text{(S4)}$$

Here, $\mathbf{R}_\mu$ and $\mathbf{R}_\sigma$ are 2x2 correlation matrices, each with one free parameter (ρ_μ and ρ_σ for the means and *SD*s, respectively) on the off-diagonal that represents the test-retest estimate for the $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$ parameters, respectively:

$$\mathbf{R}_\mu = \begin{pmatrix}1 & \rho_\mu\\ \rho_\mu & 1\end{pmatrix}$$
$$\mathbf{R}_\sigma = \begin{pmatrix}1 & \rho_\sigma\\ \rho_\sigma & 1\end{pmatrix} \quad \text{(S5)}$$

It is these free parameters (ρ_μ and ρ_σ) that we present as the test-retest reliability estimates in Figures 6-7 of the main text, and in Figures S2-S6 in section 6 below.

We specified prior distributions over group-level mean parameters using the following normal distributions:

$$
\begin{aligned}
\mu_{\text{mean,base},t} &\sim \mathcal{N}(0,1) \\
\sigma_{\text{mean,base},t} &\sim \mathcal{N}(0,1) \\
\mu_{\text{mean},\Delta,t} &\sim \mathcal{N}(0,1) \\
\sigma_{\text{mean},\Delta,t} &\sim \mathcal{N}(0,1)
\end{aligned} \tag{S6}
$$

Then, we specified prior distributions over group-level *SD* parameters with half-normal distributions (i.e. truncated at 0 to ensure that the *SD*s were greater than 0):

$$
\begin{aligned}
\mu_{\text{sd,base},t} &\sim \text{half-}\mathcal{N}(0,1) \\
\sigma_{\text{sd,base},t} &\sim \text{half-}\mathcal{N}(0,1) \\
\mu_{\text{sd},\Delta,t} &\sim \text{half-}\mathcal{N}(0,1) \\
\sigma_{\text{sd},\Delta,t} &\sim \text{half-}\mathcal{N}(0,1)
\end{aligned} \tag{S7}
$$

Lastly, we specified the priors on the correlation matrices using the LKJ distribution, which is a multivariate extension of the beta distribution that places probability density between -1 and 1 (the appropriate bounds for a correlation):

$$
\begin{aligned}
\mathbf{R}_{\mu} &= \text{LKJcorr}(1) \\
\mathbf{R}_{\sigma} &= \text{LKJcorr}(1)
\end{aligned} \tag{S8}
$$

Note that this prior specification places a non-informative, uniform probability distribution on ρ_{μ} and ρ_{σ} between -1 and 1.

Equations S2-S8 apply to each of the three generative response time models. However, for the shifted lognormal model, we had to specify additional distributions on the shift parameter ($\delta_{i,t}$). The shift parameter is more complex than other parameters in that it has a lower bound at 0 (the response time distribution cannot be shifted below 0) and an upper bound at the minimum response time for each participant. Intuitively, if we assume that the shift parameter indeed reflects non-decision factors, and that non-decision-factors determine how rapidly a person can make a response, it follows that non-decision time must be less than the minimum observed response time for a given participant i at timepoint t, or $\min(\mathbf{RT}_{i,t})$. To ensure these criteria were met, we specified individual-level shift parameters such that $\delta'_{i,t} \sim N(\delta_{\text{mean},t}, \delta_{\text{sd},t})$, where

the resulting individual-level shift parameters were then transformed and scaled by $\delta_{i,t} = \Phi(\delta'_{i,t}) \times \min(\mathbf{RT}_{i,t})$. Here, Φ is the cumulative distribution function of the standard normal distribution, which transforms the shift parameters to be between 0 and 1. Then, the parameters are scaled by the minimum response time for the corresponding participant and timepoint, ensuring that the shift parameter for each participant/timepoint is less than their fastest response time. Finally, we specified the priors on the group-level shift parameter means and *SD*s as $\delta_{\text{mean},t} \sim N(0,1)$ and $\delta_{\text{sd},t} \sim N(0,1)$, respectively.

Altogether, the prior distributions we used provide some minor-to-moderate regularization through the group-level *SD*s (i.e., greater pooling of individual-level parameters toward group-level means), but are overall weakly informative with respect to the data. This weak informativeness is apparent in the posterior predictive simulations presented for each model and task in Figures S2-S6 in section 6 below, where there is no apparent bias at the individual level (i.e. the observed response time distributions are re-produced quite well, with no obvious biases apart from model misfit). See also our parameter recovery study in section 4 below.

3.2 Delay Discounting Model

For the delay discounting task, we fit the following hyperbolic discounting function to each participant's trial-level choices:

$$V = \frac{A}{1 + kt} \tag{S9}$$

Here, V is the subjective value of the delay reward, A is the objective reward amount being discounted, and k is a discounting rate that captures how strongly rewards are discounted with increasingly long time delays t. V is computed for each of the two choices available on the current trial (e.g., "*$10 Now or $20 in 1 week?*"), and the resulting subjective values for the

smaller-sooner (V_{SS}) and larger-later (V_{LL}) choices are entered into a logistic function to determine probabilities of choosing larger-later choices:

$$\Pr(\text{choose } LL) = \frac{1}{1 + \exp(-c \cdot [V_{LL} - V_{SS}])} \tag{S10}$$

Above, c is a choice sensitivity parameter, analogous to a dispersion parameter in response time models in that higher values lead to less consistent choices (i.e., "dispersed") with respect to the discounted value for each choice option. As c increases, participants increasingly choose options with higher subjective value. Conversely, as c decreases toward 0, participants become indifferent between options, regardless of subjective values.

We used the same general group-level parameterizations to estimate test-retest reliability of delay discounting model parameters as in the response time models. The biggest difference is that we estimated test-retest correlations between the log discounting rate (k) and choices sensitivity (c) parameters, as opposed to between the raw $k_{i,t}$ and $c_{i,t}$ parameters:

$$\begin{aligned}
\begin{bmatrix}\log(k_{i,1})\\ \log(k_{i,2})\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}k_{\text{mean},1}\\ k_{\text{mean},2}\end{bmatrix}, \mathbf{S}_k\right)\\
\begin{bmatrix}\log(c_{i,1})\\ \log(c_{i,2})\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}c_{\text{mean},1}\\ c_{\text{mean},2}\end{bmatrix}, \mathbf{S}_c\right)
\end{aligned} \tag{S11}$$

As in the response time models, the covariance matrices for each parameter are again decomposed into the group-level parameter variances and correlation matrices:

$$\begin{aligned}
\mathbf{S}_k &= \begin{pmatrix}k_{\text{sd},1} & 0\\ 0 & k_{\text{sd},2}\end{pmatrix} \mathbf{R}_k \begin{pmatrix}k_{\text{sd},1} & 0\\ 0 & k_{\text{sd},2}\end{pmatrix}\\
\mathbf{S}_c &= \begin{pmatrix}c_{\text{sd},1} & 0\\ 0 & c_{\text{sd},2}\end{pmatrix} \mathbf{R}_c \begin{pmatrix}c_{\text{sd},1} & 0\\ 0 & c_{\text{sd},2}\end{pmatrix}
\end{aligned} \tag{S12}$$

Here, $\mathbf{R}_k$ and $\mathbf{R}_c$ are 2x2 correlation matrices, each with one free parameter (ρ_k and ρ_c) on the off-diagonal that represents the test-retest estimate for the $\log(k_{i,t})$ and $\log(c_{i,t})$ parameters, respectively:

$$\mathbf{R}_k = \begin{pmatrix} 1 & \rho_k \\ \rho_k & 1 \end{pmatrix}$$
$$\mathbf{R}_c = \begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix} \tag{S13}$$

For the prior distributions on the group-level mean discounting rates, we specified the following normal distributions:

$$k_{\text{mean},t} \sim \mathcal{N}(0, 1)$$
$$c_{\text{mean},t} \sim \mathcal{N}(0, 1) \tag{S14}$$

Next, we specified the group-level standard deviations using half-normal distributions truncated below at 0:

$$k_{\text{sd},t} \sim \text{half-}\mathcal{N}(0, 0.2)$$
$$c_{\text{sd},t} \sim \text{half-}\mathcal{N}(0, 0.2) \tag{S15}$$

Lastly, we used the same LKJ prior distributions (with shape parameter 1) on the correlation matrices $\mathbf{R}_k$ and $\mathbf{R}_c$ as we did for the response time models (see Equation S8).

Similar to the response time models, the prior distributions we used for the delay discounting model produce minor-to-moderate regularization through the group-level standard deviations, but the priors are otherwise relatively uninformative with respect to the data. The posterior predictive distributions in Figure S6 provides some evidence that our prior specifications did not bias individual-parameter estimates in any way that would compromise performance (see also our parameter recovery analysis in Figure S2).

3.3 Group-level Re-parameterizations to Increase Efficiency

To make the MCMC sampling more efficient, we used non-centered parameterizations for the hierarchical (group-level) components of the model, along with Cholesky decompositions to re-parameterize the test-retest correlation matrices and make sampling from the multivariate

normal distribution more efficient. Note that these re-parameterizations do not change the interpretation of the model or resulting parameter estimates—the underlying mathematical model is identical. Instead, they transform the parameters in ways that make the joint posterior distribution easier to explore from a computational perspective.

Using the individual-level mean parameters for the congruent condition in the response time models ($\mu_{i,base,t}$) as an example, the non-centered parameterization was as follows:

$$\begin{aligned} \mu_{\text{mean,base},t} &\sim \mathcal{N}(0,1) \\ \mu_{\text{sd,base},t} &\sim \text{half-}\mathcal{N}(0,1) \\ \mu'_{i,\text{base},t} &\sim \mathcal{N}(0,1) \\ \mu_{i,\text{base},t} &= \mu_{\text{mean,base},t} + \mu_{\text{sd,base},t} \times \mu'_{i,\text{base},t} \end{aligned} \tag{S16}$$

Note that this parameterization is mathematically identical to the version presented above, but now the dependency between the group-level mean, standard deviation, and individual-level parameters in the joint posterior is reduced by sampling the individual-level parameters independently from the group-level parameters, which leads to more well-behaved, elliptical bivariate distributions between the group-level means, standard deviations, and individual-level parameters (Betancourt & Girolami, 2013). We used the same general non-centered scheme for the individual-level standard deviation parameters in the base condition (where c = 1) across all response time models. Additionally, we used this parameterization for the shift parameters in the shifted lognormal model.

For individual-level parameters drawn from multivariate normal distributions, which we used to estimate test-retest correlations, we used a Cholesky decomposition to employ non-centered parameterizations. In particular, a covariance matrix $\mathbf{S}$ can be decomposed into its Cholesky factor $L_{\mathbf{S}}$, where $\mathbf{S} = L_{\mathbf{S}}L_{\mathbf{S}}^{T}$. Then, individual-level parameters drawn from independent, standard normal distributions can be correlated using the Cholesky factor $L_{\mathbf{S}}$, as shown below in Equation

S17. Importantly, the Cholesky factor ($L_{\mathbf{S}}$) of the covariance matrix **S** is equal to the diagonal matrix of the group-level *SD*s multiplied by the Cholesky factor of the correlation matrix $L_{\mathbf{R}}$. Therefore, as opposed to sampling directly from a multivariate normal distribution to estimate a correlation matrix **R**, we can sample from the group-level means, standard deviations, the Cholesky factor of the correlation matrix ($L_{\mathbf{R}}$), and individual-level parameters ($\mu'_{i,\Delta,1}$ and $\mu'_{i,\Delta,2}$) independently and then reconstruct the correlation matrix afterward as $\mathbf{R} = L_{\mathbf{R}} L_{\mathbf{R}}^T$. For example, we used the following parameterization for the individual-level mean change scores ($\mu_{i,\Delta,t}$) in the response time models:

$$
\begin{aligned}
\mu_{\text{mean},\Delta,t} &\sim \mathcal{N}(0,1) \\
\mu_{\text{sd},\Delta,t} &\sim \text{half-}\mathcal{N}(0,1) \\
\mu'_{i,\Delta,1}, \mu'_{i,\Delta,2} &\sim \mathcal{N}(0,1) \\
L_{\mathbf{R}_\mu} &\sim \text{LKJcorr}(1) \\
L_{\mathbf{S}_\mu} &= \begin{pmatrix} \mu_{\text{sd},\Delta,1} & 0 \\ 0 & \mu_{\text{sd},\Delta,2} \end{pmatrix} L_{\mathbf{R}_\mu} \\
\begin{bmatrix} \mu_{i,\Delta,1} \\ \mu_{i,\Delta,2} \end{bmatrix} &= \begin{bmatrix} \mu_{\text{mean},\Delta,1} \\ \mu_{\text{mean},\Delta,2} \end{bmatrix} + L_{\mathbf{S}_\mu} \begin{bmatrix} \mu'_{i,\Delta,1} \\ \mu'_{i,\Delta,2} \end{bmatrix} \\
\mathbf{R}_\mu &= L_{\mathbf{R}_\mu} L_{\mathbf{R}_\mu}^T
\end{aligned}
\tag{S17}
$$

Despite the multivariate normal distribution not being explicitly defined above, the individual-level parameters have the relationship such that $\mu_{i,\Delta,1}$ and $\mu_{i,\Delta,2}$ follow a multivariate normal distribution, with means $\mu_{\text{mean},\Delta,1}$ and $\mu_{\text{mean},\Delta,2}$ and covariance matrix $\mathbf{S}_\mu = L_{\mathbf{S}_\mu} L_{\mathbf{S}_\mu}^T$. Like the non-centered parameterization for independent individual-level parameters shown in Equation S16, this non-centering reduces dependence between the group-level means, standard deviations, correlations, and individual-level parameters in the joint posterior distribution. The Cholesky decomposition also has the added benefit of speeding up computation by side-stepping the need to invert the covariance matrix $\mathbf{S}_\mu$ when evaluating the multivariate normal distribution during MCMC sampling.

We used the same general re-parameterization scheme described in Equation S17 to parameterize $\log(k_{i,t})$ and $\log(c_{i,t})$ in the delay discounting model (but with the group-level priors specified as in Equations S14 and S15). We refer readers to the Stan user's manual for more information, which has an excellent section on the non-centered and Cholesky factorization parameterizations employed above (https://mc-stan.org/docs/2_23/stan-users-guide/reparameterization-section.html).

3.4 Parameter Estimation

For each of the response time models and for the delay discounting model, we fit three separate sampling chains, each for 3,000 samples, wherein the first 1,000 were used and discarded for warm-up/tuning. Sampling therefore resulted in a total of 6,000 posterior samples for each parameter. We visually checked for convergence to the target distribution using traceplots, and we also ensured that all Gelman-Rubin diagnostics ($\hat{R}$) were below 1.1 (Gelman & Rubin, 1992).

4. Parameter Recovery Simulation Study

We conducted parameter recovery simulations to determine how well the proposed generative models could recover known test-retest correlations. To do so, we simulated response times from the full lognormal generative model, where we set the group-level mean and standard deviation parameters to the following specific values:

$$
\begin{aligned}
\mu_{\text{mean},base,1} &= -0.5 \\
\mu_{\text{mean},base,2} &= -0.55 \\
\mu_{\text{mean},\Delta,1} &= 0.1 \\
\mu_{\text{mean},\Delta,2} &= 0.08 \\
\mu_{\text{sd},base,1} &= 0.11 \\
\mu_{\text{sd},base,2} &= 0.12 \\
\mu_{\text{sd},\Delta,1} &= 0.035 \\
\mu_{\text{sd},\Delta,2} &= 0.029 \\
\\
\sigma_{\text{mean},base,1} &= -1.28 \\
\sigma_{\text{mean},base,2} &= -1.32 \\
\sigma_{\text{mean},\Delta,1} &= 0.12 \\
\sigma_{\text{mean},\Delta,2} &= 0.11 \\
\sigma_{\text{sd},base,1} &= 0.19 \\
\sigma_{\text{sd},base,2} &= 0.2 \\
\sigma_{\text{sd},\Delta,1} &= 0.11 \\
\sigma_{\text{sd},\Delta,2} &= 0.09
\end{aligned}
\tag{S18}
$$

Holding the parameters in equation S18 constant, we then varied the true test-retest correlations (ρ_μ and ρ_σ in equation S5) across a grid of 15 evenly spaced values ranging from -1 to 1 (ρ_μ and ρ_σ shared the same correlation at each position in the grid). For each true correlation in the grid, the simulation proceeded by first generating individual-level parameters from the group-level distributions. Next, response times within each task condition were simulated from the individual-level distributions using parameters simulated in the previous step. In addition to varying the true correlation, we varied the number of simulated participants and trials within each condition to determine the effect of sample size on parameter recovery. Specifically, for each true correlation, we simulated: (1) 10 participants with 10 trials per condition, (2) 50 participants with 50 trials per condition, (3) 50 participants with 100 trials per condition, (4) 100 participants with 50 trials per condition, and (5) 100 participants with 100 trials per condition. We chose these sample sizes as they are typical in many studies using behavioral tasks (apart

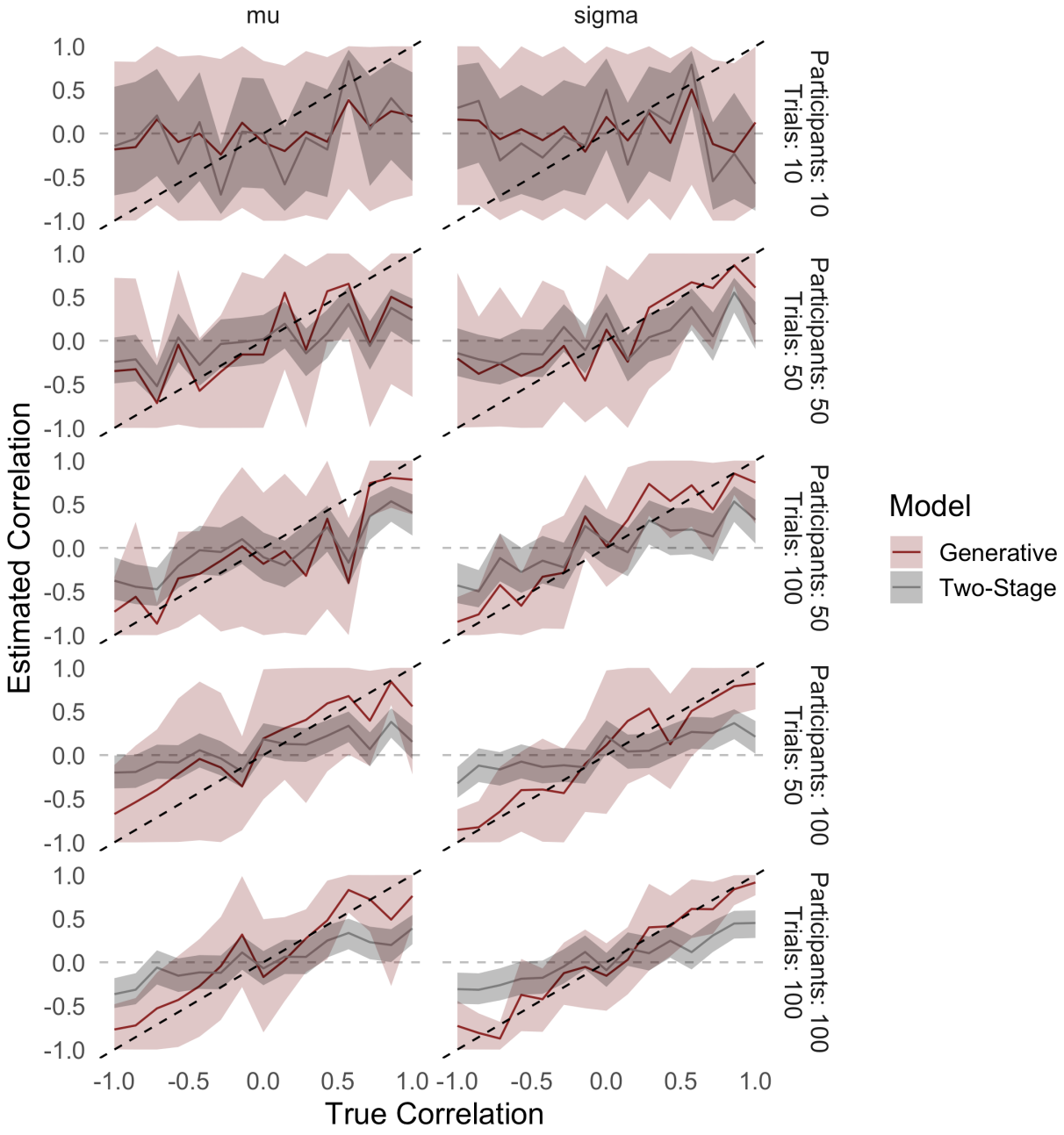
from the 10 participant/trial example, which we included to demonstrate how a lack of information influences generative model estimates).

After simulating the response times from the full lognormal generative model, for each correlation in the grid and for each participant/trial configuration we fit both: (1) the two-stage model (i.e. computing means and standard deviations for each participant/condition/session, computing condition contrasts, and then estimating test-retest correlations as the correlation across sessions), and (2) the full lognormal generative model with the same specification as described in the main text and above (see section 3 above). The parameter recovery results are presented in Figure S2, which shows how well each approach can recover the true generating correlation. As is clear in the figure, the two-stage approach does a poor job recovering the true correlation—the correlation estimates are attenuated toward 0 (i.e. regressing toward 0 due to low reliability), and the 95% confidence intervals are much narrower than they should be (due to the assumption that individual-level mean/standard deviation contrasts are estimated with no measurement error). In contrast, the generative model performs quite well—it exhibits good coverage properties, where the 95% highest density intervals almost always contain the true underlying correlation. Further, as shown in the low information condition (10 participants with 10 trials per condition), the generative model appropriately calibrates our uncertainty and does not exhibit systematic biases in low data settings (e.g., from our choice of prior distributions). Note that the generative model continues to outperform the two-stage approach even in relatively high information settings—with 100 participants and 100 trials per condition, the expected correlation inferred from the generative model closely recovers the true correlations across the entire grid. In contrast, the two-stage approach continues to underestimate both the magnitude

and uncertainty associated with the correlation, particularly when the true correlation is of large magnitude (i.e. close to -1 or 1).

There are two important take-aways from our parameter recovery study. First, we do not need to worry about the reliability of either measure when fitting a full generative model—the posterior distribution will always reflect the appropriate level of uncertainty given our model, assumed prior distributions, and observed data. This can be observed in the low data condition (top panel of Figure S2), where the posterior distribution covers practically the entire range of possible correlations from -1 to +1. In this case, the posterior distribution indicates that we have learned almost nothing about the test-retest correlation from our data, and that we would need more data to make a more precise inference. Second, our results make it clear that estimates derived from the two-stage approach cannot be trusted as-is. One could use post-hoc corrections for attenuation due to low reliability to better estimate the true correlation (e.g., Spearman, 1904), although this procedure does not come along with an agreed upon method for correcting resulting confidence intervals. Further, such techniques require us to already know the reliability of our measure (e.g., the mean contrast), which necessitates either more data collection or knowledge of the sampling distribution underlying our individual-level estimates.

Figure S2. Generative versus two-stage approach to recovering a true test-retest correlation. "mu" and "sigma" indicate the correlations for the mean and standard deviation parameters, respectively. Intervals are 95% highest density intervals and 95% CIs for the generative and two-stage estimates, respectively. The diagonal dotted line represents the true correlation around which the uncertainty intervals should cover to indicate successful parameter recovery.



5. Detailed Description of Datasets and Behavioral Paradigms

5.1 Attention and Inhibitory Control

We include response time data for the Stroop, Flanker, and Posner Cueing tasks from Hedge et al. (2017). For the Stroop and Flanker tasks, two sets of participants (n = 47, n = 60 for Studies 1 and 2, as reported in the original work) performed each task twice, separated by three weeks. For the Posner Cueing task, a third set of participants (n = 40 for Study 3 in the original work) performed the task twice, also separated by three weeks. We include more participants than in the original analyses because we did not use heuristic preprocessing rules (Hedge et al., 2017).

All three tasks have a similar structure, whereby mains effects of interest are contrasts between congruent and incongruent conditions. In the Stroop task, participants responded to the color of a word, which could be red, blue, green, or yellow. The word could be the same as the font color (e.g., the word "red" colored in red font; congruent condition or c = 1), a non-color word (e.g., "ship"; neutral condition), or a color word mapping onto another response option (e.g., the word "red" colored blue, green, or yellow; incongruent condition or c = 2). The Flanker task is similar. Participants respond to the direction of a central arrow, which is surrounded by arrows in the same direction (congruent condition or c = 1), straight lines (neutral condition), or arrows in the opposite direction (incongruent condition or c = 2). In tasks, participants completed 240 trials in each of the three conditions.

For the Posner Cueing task, participants respond to a stimulus ("X"), which can appear in one of two boxes on the left or right side of a central fixation. Before the X stimulus appears, an arrow (cue) points to either the left or right box. On 80% of trials, the direction of the cue predicts the location of the subsequent X stimulus (congruent condition or c = 1), and on the

remaining trials the X stimulus appears in the opposite box (incongruent condition or $c = 2$). Participants completed 640 trials total.

5.2 Implicit Associations

We included Self-Concept (introversion/extraversion) and Race (Black/White) versions of the Implicit Association Test (IAT) using data from Gawronski et al. (2017). For the Self-Concept IAT (Study 1a of the original work), 152 participants completed the task twice, separated by approximately eight weeks. For the Race IAT, (Study 2b of the original work), 116 participants completed the task twice, also separated by approximately eight weeks.

The Self-Concept IAT comprised five blocks, two of which are of interest here. In Block 3 ($c = 1$), participants classified self-related words (e.g., *I*, *me*, *mine*) and extraversion-related words (e.g., *active*, *talkative*, *sociable*) with one response option (right-hand key labelled *Me*), and non-self-related words (e.g., *few*, *some*, *any*) and introversion words (e.g., *passive*, *quiet*, *withdrawn*) with the alternative response option (left-hand key labelled *Not Me*). In Block 5 ($c = 2$), mapping was reversed such that participants classified self- and introversion-related words with the right-hand key and non-self- and extraversion-related words with the left-hand key. Participants completed 80 trials within each block.

The Race IAT followed the same structure. In Block 3 ($c = 1$), participants classified images of White individual's faces and positive words (e.g., *good*, *pleasant*, *likable*) with a right-hand key, and images of Black individual's faces and negative words (e.g., *bad*, *unpleasant*, *dislikable*) with a left-hand key. In Block 5 ($c = 2$), mapping was reversed such that participants classified images of Black individual's faces and positive words with a right-hand key and

images of White individual's faces and negative words with a left-hand key. Participants completed 60 trials within each block.

5.3 Impulsivity

We include the staircase procedure variant of delay discounting task data collected by Ahn et al. (2020). Participants (n = 58) completed the delay discounting task four separate times: twice during each of two visits, with visits separated by four weeks. We include data from the first administration of the task within each visit.

During each task administration, participants completed 42 trials in which they made preference judgements between smaller-sooner and larger-later choices (e.g., *would you rather have $10 Now [smaller-sooner] or $20 in 1 week [larger-later]?*). The staircase procedure starts with the choice between $400 now or $800 at 1 of 7 delays: 1 week, 2weeks, 1 month, 6 months, 1 year, 3 years, or 10 years. After each choice, the smaller-sooner amount is adjusted by 50% of the preceding increment (starting with a $200 increment) in a direction to increase the subjective value of the unchosen option. This procedure is iterated until the increment reaches $12.50 (for further details see Ahn et al., 2020). The indifference point for each of the seven delays is set as the minimum dollar amount of the smaller-sooner option that participants choose within the given delay. If participants only choose the larger-later option within a delay, the indifference point is set to $800 (the maximum dollar amount across options), indicating no discounting of the larger-later reward.

6. Detailed Empirical Results

For each task/study below, we include high-level descriptions of results along with visual depictions of test-retest estimates and posterior predictive simulations. We use the term "expected" to refer to the posterior mean estimates of the generative model parameters. Table S1 contains point estimates and uncertainty intervals for test-retest estimates for each model, including tasks and parameters (see Figure 6 in the main text for a visual representation).

Table 1. Test-retest results for all tasks and models

Task/Study	Model	Parameter	Estimate	95% Interval
Stroop Study 1	Two-stage Approach	Sample Mean	.50	[.25, .69]
		Sample SD	.07	[-.22, .35]
	Normal	μ_Δ	.76	[.46, 1.00]
		σ_Δ	.23	[-.06, .50]
	Lognormal	μ_Δ	.77	[.47. 1.00]
		σ_Δ	.60	[.26, .89]
	Shifted-Lognormal	μ_Δ	.81	[.53, 1.00]
		σ_Δ	.62	[.25, .96]
Stroop Study 2	Two-stage Approach	Sample Mean	.63	[.45, .76]
		Sample SD	.34	[.10, .55]
	Normal	μ_Δ	.84	[.67, .98]
		σ_Δ	.37	[.15, .60]
	Lognormal	μ_Δ	.82	[.65, 1.00]
		σ_Δ	.48	[.16, .76]
	Shifted-Lognormal	μ_Δ	.75	[.53, .93]
		σ_Δ	.54	[.15, .91]
Flanker Study 1	Two-stage Approach	Sample Mean	.32	[.03, .55]
		Sample SD	-.02	[-.31, .26]
	Normal	μ_Δ	.71	[.38, 1.00]
		σ_Δ	-.03	[-.33, .25]
	Lognormal	μ_Δ	.73	[.42, 1.00]
		σ_Δ	.11	[-.19, .41]
	Shifted-Lognormal	μ_Δ	.71	[.44, .95]
		σ_Δ	.14	[-.18, 47]
Flanker Study 2	Two-stage Approach	Sample Mean	-.13	[-.37, .13]
		Sample SD	.12	[-.14, .36]
	Normal	μ_Δ	.64	[.35, .89]
		σ_Δ	.09	[-.16, 35]
	Lognormal	μ_Δ	.73	[.48, .96]

	Shifted-Lognormal	σ_Δ	.07	[-.22, .37]
		μ_Δ	.74	[.54, .92]
		σ_Δ	.20	[-.13, .51]
Posner Study 3	Two-stage Approach	Sample Mean	.17	[-.15, .46]
		Sample SD	.21	[-.11, .49]
	Normal	μ_Δ	.78	[.55, .98]
		σ_Δ	-.06	[-.39, .26]
	Lognormal	μ_Δ	.81	[.54, 1.00]
		σ_Δ	-.03	[-.36, .31]
	Shifted-Lognormal	μ_Δ	.80	[.52, 1.00]
		σ_Δ	-.01	[-.35, .32]
IAT Self-Concept	Two-stage Approach	Sample Mean	.60	[.49, .69]
		Sample SD	.39	[.25, .52]
	Normal	μ_Δ	.73	[.63, .82]
		σ_Δ	.53	[.42, .65]
	Lognormal	μ_Δ	.69	[.59, .78]
		σ_Δ	.60	[.47, .71]
	Shifted-Lognormal	μ_Δ	.67	[.56, .76]
		σ_Δ	.40	[.21, .58]
IAT Race	Two-stage Approach	Sample Mean	.45	[.30, .59]
		Sample SD	.15	[-.03, .32]
	Normal	μ_Δ	.83	[.73, .93]
		σ_Δ	.32	[.15, .50]
	Lognormal	μ_Δ	.63	[.47, .78]
		σ_Δ	.39	[.19, .58]
	Shifted-Lognormal	μ_Δ	.57	[.42, .74]
		σ_Δ	.37	[.14, .57]
Delay Discounting	Two-stage MLE with Hyperbolic Model	k	.64	[.46, .77]
		c	.54	[.33, .70]
	Hierarchical Bayesian with Hyperbolic Model	k	.74	[.63, .84]
		c	.73	[.55, .90]

Note. This table contains descriptions of the test-retest correlations for all the tasks analyzed in the current study. 95% intervals indicate the 95% highest density interval for generative models, and the 95% confidence interval for traditional two-stage summary statistic or MLE (maximum

likelihood estimation) approaches. The test-retest windows for each task were approximately as follows: (1) 3 weeks for the Stroop, Flanker, and Posner Cueing tasks, (2) 8 weeks for both versions of the IAT, and (3) 4 weeks for the Delay Discounting task.

Additionally, a key component of generative modeling is identifying areas of model misfit that could (1) influence how we interpret the results, and (2) offer insight into potential extensions of our models. Given our interest in individual-differences within each behavioral paradigm, we used visual checks to determine whether each model provided adequate fit to the observed data at the individual participant level. We simulated data from each participant's individual-level parameter estimates and examined how well the simulations matched observed behavior across task conditions. In Bayesian terminology, these simulations are referred to as *posterior predictive simulations*. Posterior predictive simulations for response times models involve simulating response times from each of the normal, lognormal, and shifted lognormal models and checking the extent to which simulated response time distributions are similar to observed response time distributions within each condition. For the hyperbolic model, we plotted estimated discounting curves against participants' empirical indifference points (i.e., the point at which they become indifferent to the smaller-sooner and larger-later reward using a staircase procedure). Posterior predictive simulations for each task and model are presented along with the results in the following section.

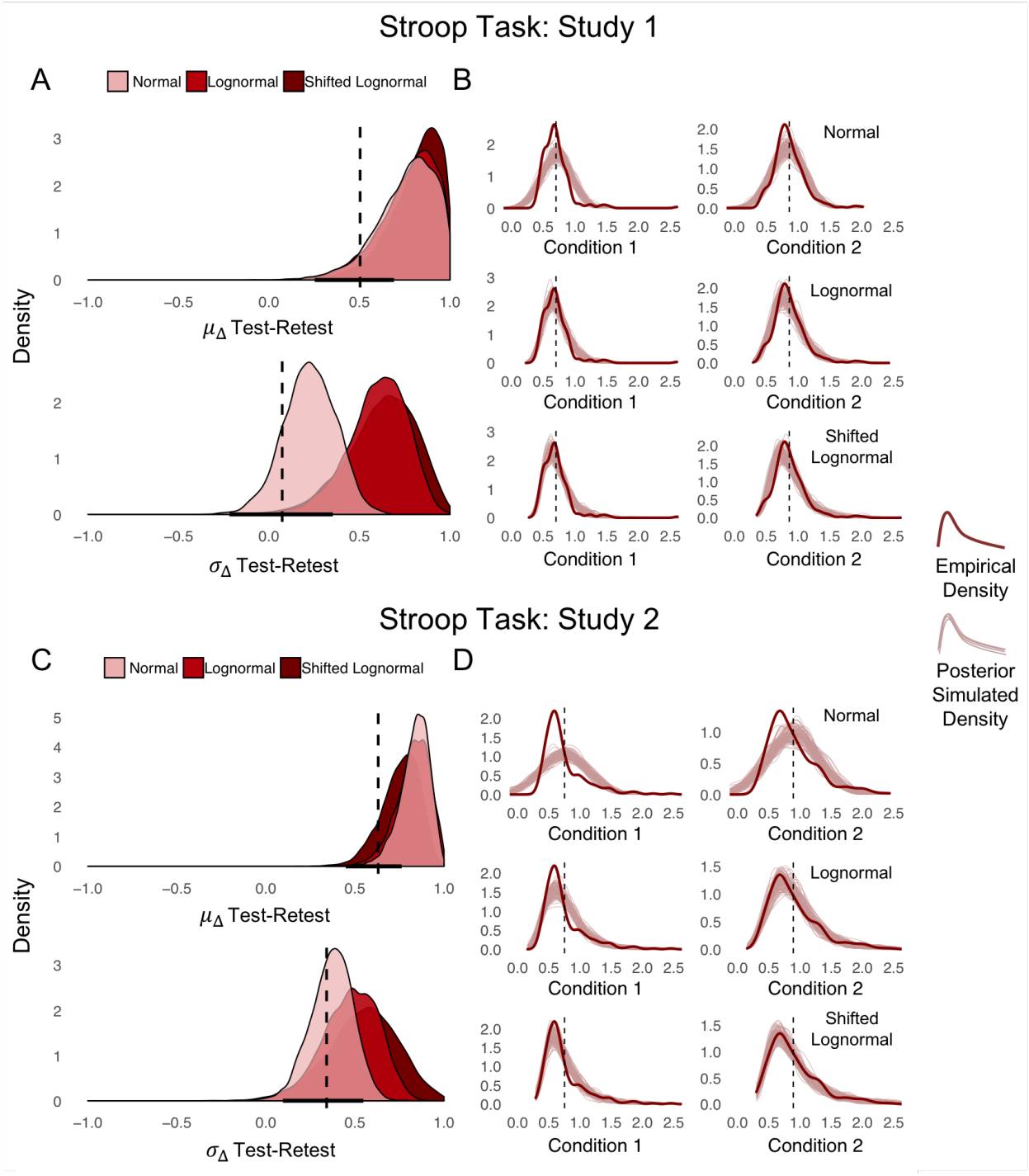
6.1 Stroop Task

Test-retest results for the response time models applied to the Stroop task were much different for the two-stage approach compared to the generative models (see Figure S2). Using the two-stage sample mean/standard deviation approach, test-retest correlations between the mean contrasts were $r = .50$ and $r = .63$ for Hedge et al.'s (2017) Studies 1 and 2, respectively. The standard deviation contrasts were much lower, with test-retest correlations of $r = .07$ and $r = .34$. For the generative models across both studies, the posterior distributions for the

mean/difficulty parameters ($\mu_{i,\Delta}$) were concentrated above the two-stage estimates (expected test-retest ranging from r = .75 to r = .84). Posterior distributions for the dispersion parameters ($\sigma_{i,\Delta}$) were also concentrated above the two-stage estimates, although primarily for the lognormal and shifted lognormal models (expected test-retest ranging from r = .23 to r = .62).

For all three generative models and each study, Figure S3 also shows the posterior predictive simulations for a random, representative participant. It is clear that the normal generative model—and by extension, the two-stage approach—does a poor job of capturing the strong right skew from response time distributions. In contrast, the lognormal and shifted lognormal models perform well, yielding an increase in expected test-retest reliability for the dispersion parameters in the lognormal models over the normal model.

Figure S3. Test-retest correlations and model misfit for the Stroop task. (A) Posterior distributions for the test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the Stroop task in Study 1 of Hedge et al. (2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative subject. (C) and (D) show similar results, but for Study 2.

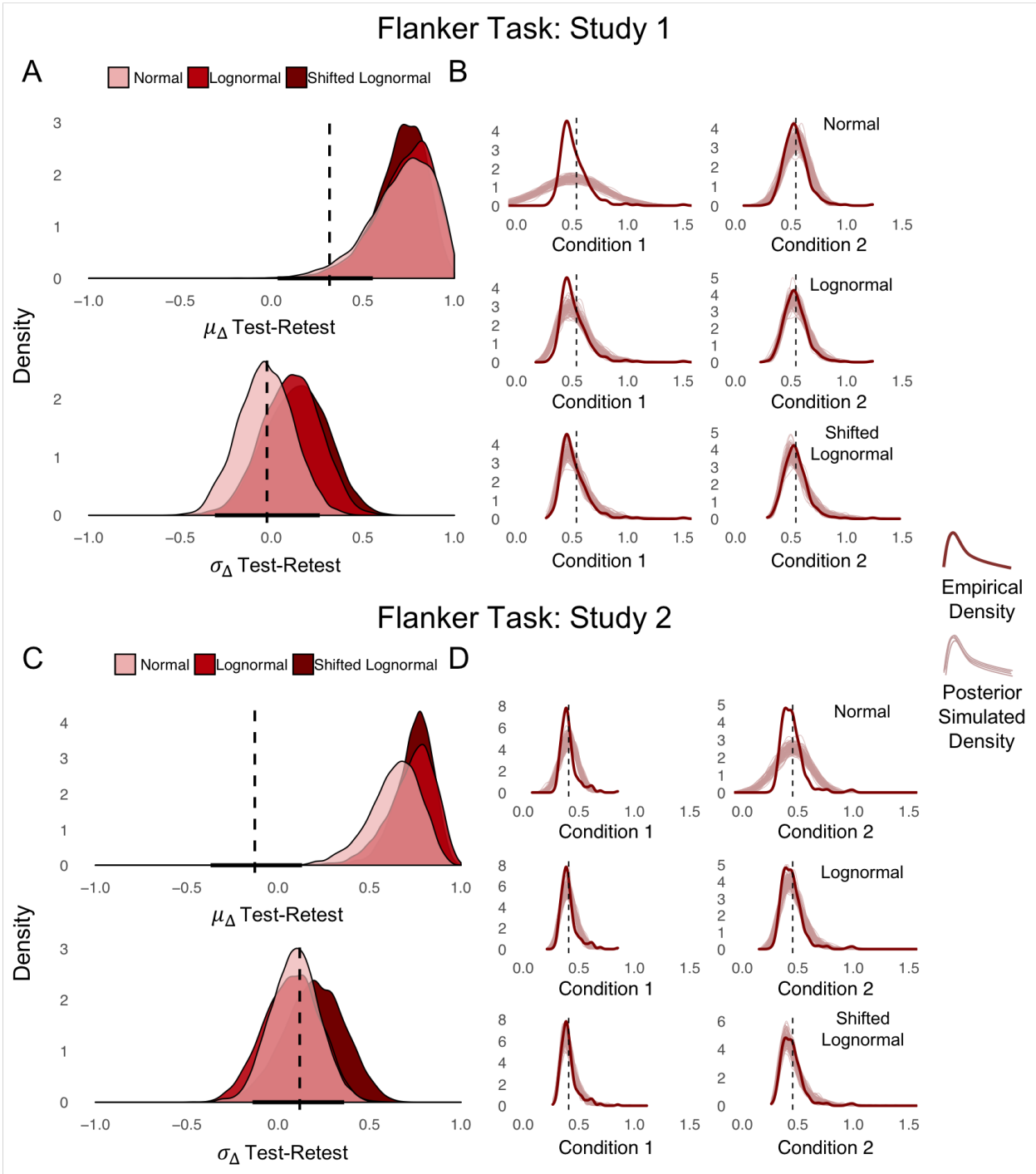

Stroop Task: Study 1
A
Normal
Lognormal
Shifted Lognormal
B
Density
μ_Δ Test-Retest
σ_Δ Test-Retest
Condition 1
Condition 2
Normal
Lognormal
Shifted Lognormal
Empirical Density
Posterior Simulated Density
Stroop Task: Study 2
C
D

6.2 Flanker Task

The results for the Flanker task were quite varied. As shown in Figure S4A and S4C, the two-stage approach produced surprisingly low test-retest estimates for both mean contrasts (r = .32 and r = -.13 for Studies 1 and 2, respectively) and standard deviation contrasts (r = -.02 and r = .12). For the generative models, however, expected test-retest estimates for the mean/difficulty parameters ($\mu_{i,\Delta}$) ranged from r = .64 to r = .74 across models and studies. Nevertheless, the expected test-retest estimates for dispersion parameters ($\sigma_{i,\Delta}$) were consistent with the two-stage approach, ranging from r = -.03 to r = .20. These results indicate that dispersion shows little to no reliable between-participant variability under the assumption of normal, lognormal, and shifted lognormal behavioral models. Importantly, such results only apply to these three behavioral models. Other behavioral models—which may contain parameters that are interpreted similarly to dispersion—will produce different inferences (e.g., compare the normal versus lognormal model dispersion test-retest estimates for the Stroop task in Figure S3A).

The posterior predictive simulations for the Flanker task (Figure S4B and S4D) corroborate those of the Stroop task—the normal generative model shows poor fit to observed response time distributions across participants, whereas lognormal and shifted lognormal models capture the distributions well. Unlike the Stroop task findings, better fit for the lognormal and shifted lognormal models does not lead to increased test-retest estimates for the dispersion parameter.

Figure S4. Test-retest correlations and model misfit for the Flanker task. (A) Posterior distributions for test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the Flanker task in Study 1 of Hedge et al. (2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative participant. (C) and (D) present similar results, but for Study 2.

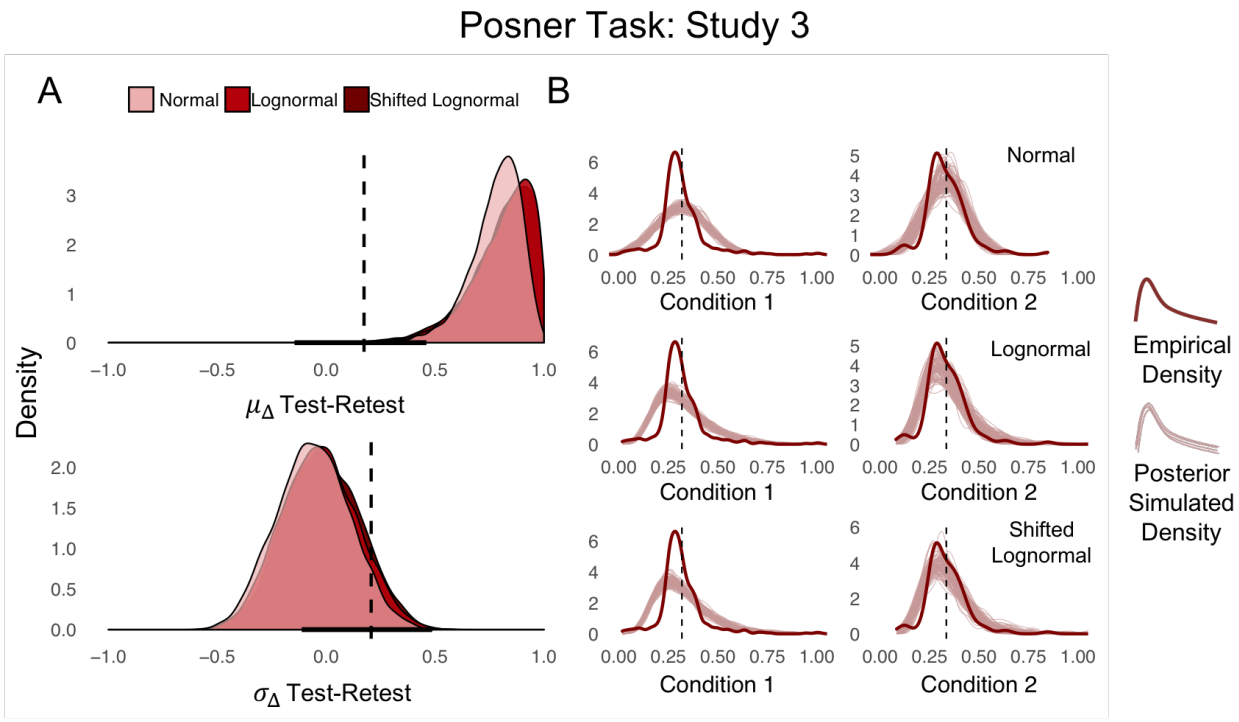

Flanker Task: Study 1
A
Normal
Lognormal
Shifted Lognormal
B
Density
μ_Δ Test-Retest
σ_Δ Test-Retest
Condition 1
Condition 2
Normal
Lognormal
Shifted Lognormal
Empirical Density
Posterior Simulated Density
Flanker Task: Study 2
C
D

6.3 Posner Task

Results from the Posner task closely mirror those from the Flanker task. The two-stage approach produced low test-retest estimates for both mean (r = .17) and standard deviation (r = .21) contrasts, and the generative models produced high expected test-retest for the mean/difficulty parameters (ranging from r = .78 to r = .81) but not for the dispersion parameters (ranging from r = -.06 to r = -.02) (Figure S5A).

The posterior predictive simulations shown in Figure S5B reveal some important discrepancies between predicted and empirical response time distributions that are not apparent in the simulations for participants in the Stroop and Flanker tasks (Figures S3 & S4). The participant depicted in Figure S5B exhibited multiple rapid response times between 0 and .2 seconds, although most of their response times fell between approximately .2-.5 seconds. Even the shifted lognormal model does a poor job capturing this participant's empirical response time distribution, indicating that caution should be taken before assigning psychological meaning to the behavioral model parameters (assuming this pattern holds across many participants). An straightforward extension to the shifted lognormal model that could resolve this problem is to model response times as arising from a mixture between the shifted lognormal process and a uniform distribution that represents "contamination" response times. Such mixture modeling is common practice in the evidence accumulation modeling literature, wherein many behavioral models (e.g., the Diffusion Decision Model, Linear Ballistic Accumulator) contain a non-decision time parameters that can be estimated poorly when contamination trials occur below what the non-decision time threshold would otherwise suggest.

Figure S5. Test-retest correlations and model misfit for the Posner Cueing task. (A) Posterior distributions for test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the Posner Cueing task in Study 3 of Hedge et al. (2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative participant.

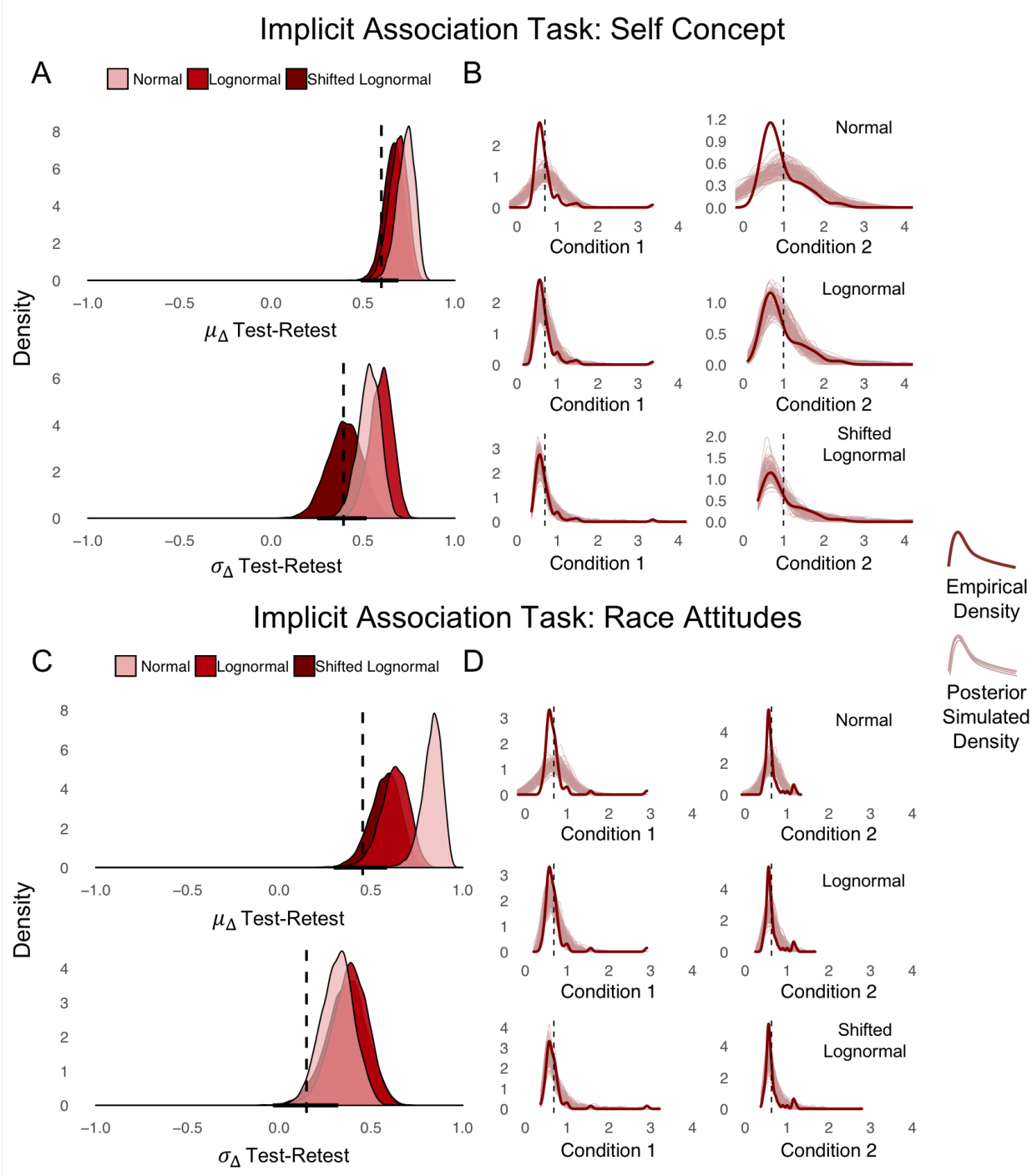


6.4 Implicit Association Test

The two-stage approach produced low to moderate test-retest correlations for mean contrasts (Self-Concept $r = .60$; Race $r = .45$) and standard deviation contrasts (Self-Concept $r = .39$; Race $r = .15$). As in the other tasks, the generative models tended to produce higher expected test-retest estimates. Across both the identity Self-Concept IAT and the Race IAT, expected mean/difficulty test-retest ranged from $r = .57$ to $r = .83$, whereas expected dispersion test-retest ranged from $r = .32$ to $r = .60$ (Figure S6A & S6C).

In general, the lognormal and shifted lognormal models provided good fit to empirical response time distributions across both versions of the Implicit Association Test. Despite lognormal models providing a much better fit to empirical response time distributions, the normal generative model yielded higher expected test-retest reliabilities for the mean/difficulty parameters (see example participants in Figure S6B & S6D).

Figure S6. Test-retest correlations and model misfit for the Implicit Association Tests. (A) Posterior distributions for the test-retest correlations of each of the three generative models (red distributions) versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) in the Self-Concept IAT from Gawronski et al. (2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative participant. (C) and (D) present similar results, but for the Race IAT.

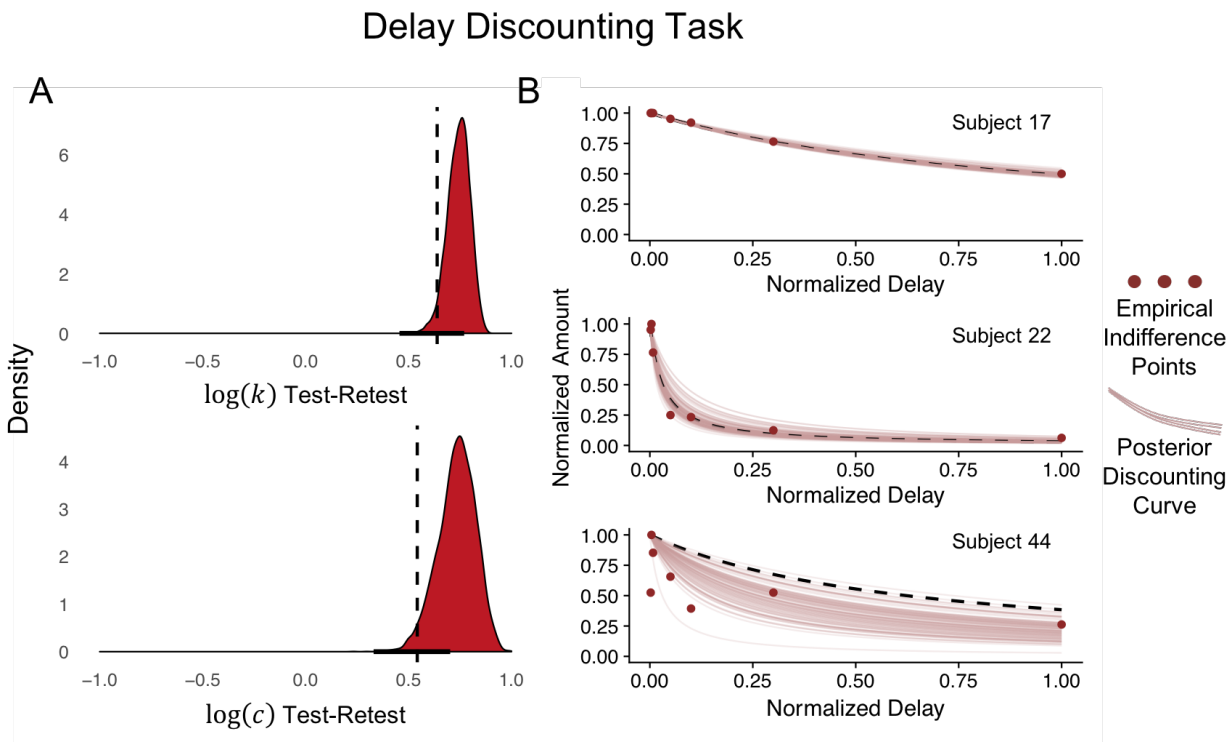

Implicit Association Task: Self Concept
A
Normal
Lognormal
Shifted Lognormal
Density
μ_Δ Test-Retest
σ_Δ Test-Retest
B
Normal
Lognormal
Shifted Lognormal
Condition 1
Condition 2
Empirical Density
Posterior Simulated Density
Implicit Association Task: Race Attitudes
C
D

6.5 Delay Discounting Task

Results for the test-retest reliability of the hyperbolic model fit to the delay discounting task show that benefits of hierarchical modeling extend beyond response time tasks. As shown in Figure S7A, the two-stage maximum likelihood approach—which is analogous to the sample mean/standard deviation approach for response time models—produced test-retest estimates of r = .64 and r = .54 for the discounting rate (k_i) and choice sensitivity (c_i) parameters, respectively (these correlations are calculated for parameters on the log scale). Similar to the response time models, the full generative model that estimated parameters hierarchically produced higher expected test-retest reliabilities for both discounting rate (expected r = .74) and choice sensitivity (expected r = .73) parameters.

To visualize the performance of both approaches at the individual level, we plotted participants' empirical indifference points against the predictions from the two-stage maximum likelihood and full generative model approaches. Figure S7B shows examples of three representative participants.

Figure S7. Test-retest correlations and model misfit for the delay discounting task. (A) Posterior distributions for the test-retest correlations of each of the three generative models (red distributions) versus the two-stage maximum likelihood estimation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the staircase version of the delay discounting task used by Ahn et al. (2020). (B) Posterior discounting curves and discounting curves estimated using maximum likelihood (dotted black lines) for three representative participants. Indifference points were computed as described in section 5.3.



7. Sensitivity Analyses

As described in the main text, we tested two different group-level models to determine how sensitive our results were to changes in generative assumptions. First, we tested an alternative group-level model wherein both the person-level base and change parameters were drawn from separate multivariate normal distributions, as opposed to only the change parameters. This change involves a modification to the group-level model specification in the main text (Equations 4 and 5) to the following:

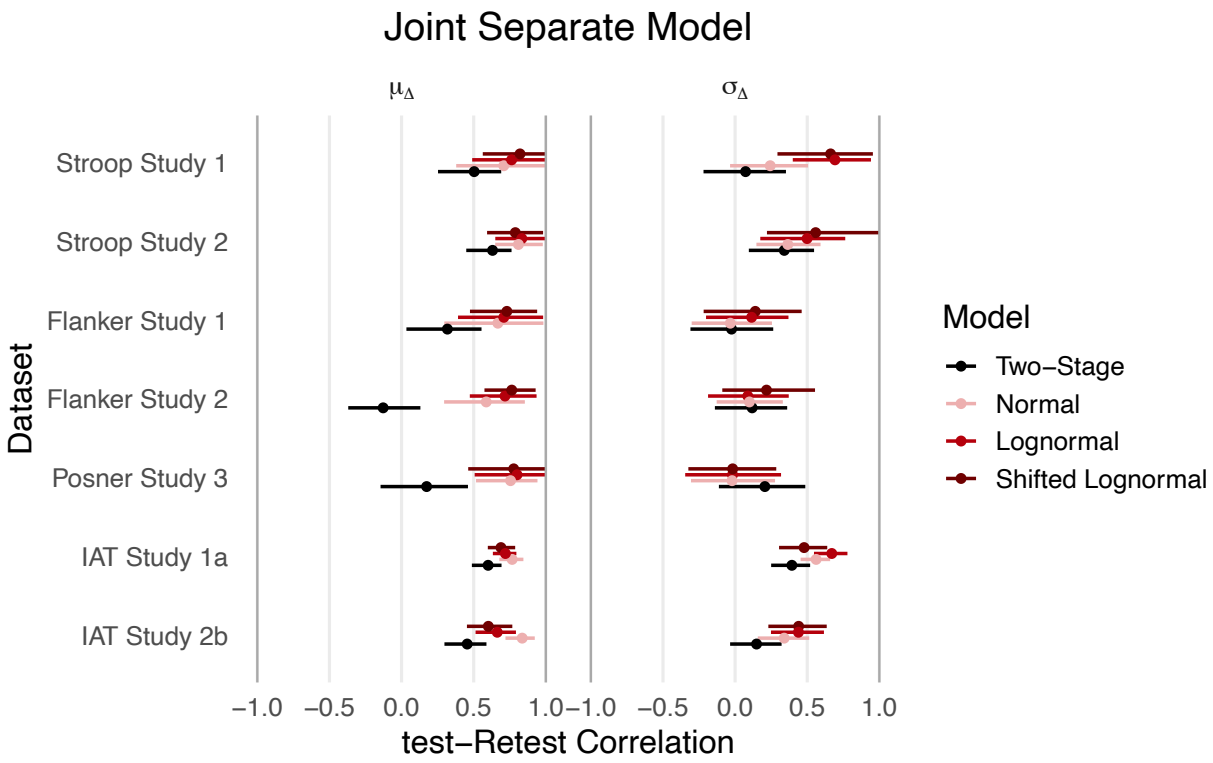
$$\begin{aligned}
\begin{bmatrix}\mu_{i,\text{base},1}\\ \mu_{i,\text{base},2}\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}\mu_{\text{mean},\text{base},1}\\ \mu_{\text{mean},\text{base},2}\end{bmatrix}, \mathbf{S}_{\mu_{\text{base}}}\right)\\
\begin{bmatrix}\sigma_{i,\text{base},1}\\ \sigma_{i,\text{base},2}\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}\sigma_{\text{mean},\text{base},1}\\ \sigma_{\text{mean},\text{base},2}\end{bmatrix}, \mathbf{S}_{\sigma_{\text{base}}}\right)\\
\\
\begin{bmatrix}\mu_{i,\Delta,1}\\ \mu_{i,\Delta,2}\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}\mu_{\text{mean},\Delta,1}\\ \mu_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\mu_{\Delta}}\right)\\
\begin{bmatrix}\sigma_{i,\Delta,1}\\ \sigma_{i,\Delta,2}\end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix}\sigma_{\text{mean},\Delta,1}\\ \sigma_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\sigma_{\Delta}}\right)
\end{aligned} \tag{S19}$$

Now, the base parameters are estimated from correlated (i.e. multivariate) normal distributions as opposed to from independent normal distributions. Note that the specification of the group-level distribution over the difference (Δ) parameters remains unchanged, although we include it here with additional subscripts on the covariance matrices ($\mathbf{S}_{\mu_\Delta}$ and $\mathbf{S}_{\sigma_\Delta}$) to differentiate them from those of the base parameters ($\mathbf{S}_{\mu_{\text{base}}}$ and $\mathbf{S}_{\sigma_{\text{base}}}$). We used the same prior distributions on the correlation matrices for the base parameters as in the difference parameters (see Equations S4-S8). We term this model the "Joint Separate" model, given that it assumes that the baseline and change parameters arise from separate group-level multivariate normal distributions. The test-retest obtained for the response time tasks are shown in Figure S8. Overall, the model produced

estimates very similar estimates to those of the model used in the main text (c.f. Figure S8 to Figure 6).

Figure S8. Test-retest correlations for all tasks using the Joint Separate models.

Here we show means and 95% confidence intervals for the two-stage summary approach (for both sample mean and standard deviation contrasts) in black, along with the posterior means and 95% highest density intervals for the generative model parameter estimates (in various shades of red). The Implicit Association Test (IAT) datasets are from the Self-Concept (introversion/extraversion; Study 1a) and Race (Black/White; Study 2b) versions. This model uses the group-level specification described in Equation S19.



Second, we tested a group-level model wherein we directly estimated the $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ parameters as opposed to estimating baseline and change parameters. For this model, we assumed that person-level parameters were drawn from a single multivariate normal distribution (but separate for μ versus σ) across conditions and sessions:

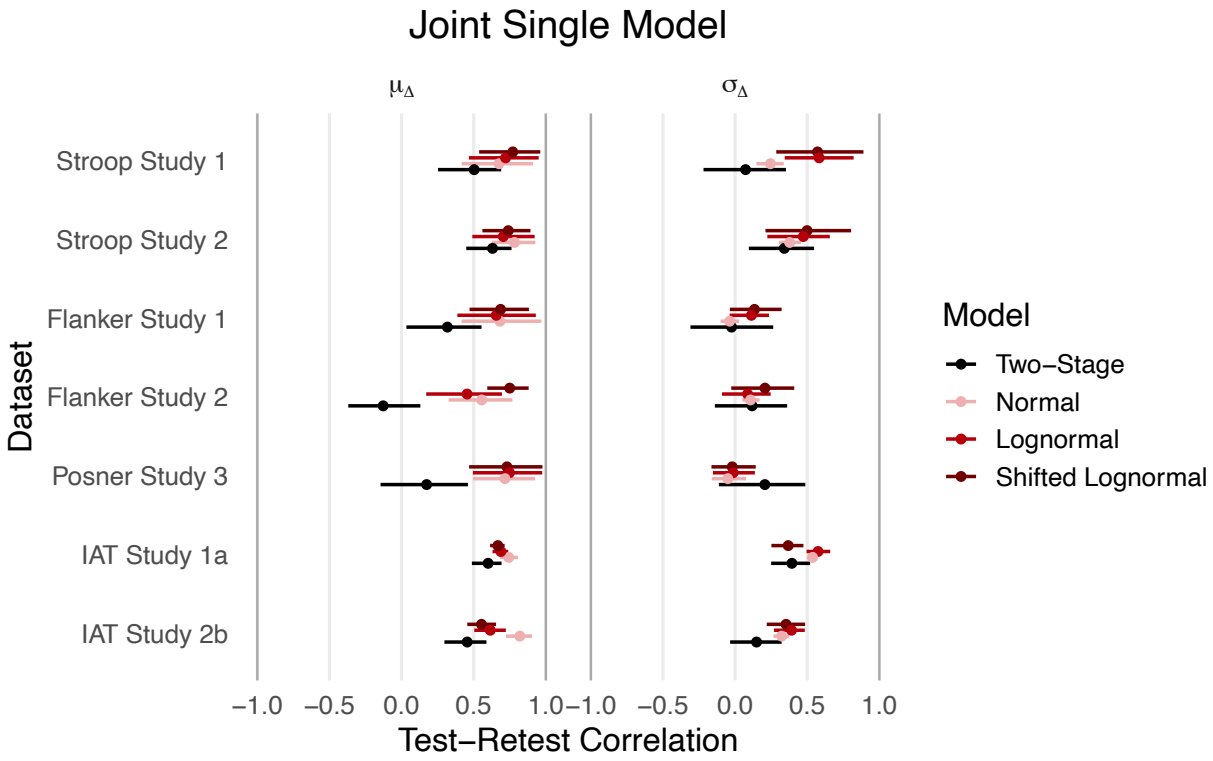
$$\begin{bmatrix} \mu_{i,1,1} \\ \mu_{i,1,2} \\ \mu_{i,2,1} \\ \mu_{i,2,2} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \mu_{\text{mean},1,1} \\ \mu_{\text{mean},1,2} \\ \mu_{\text{mean},2,1} \\ \mu_{\text{mean},2,2} \end{bmatrix}, \mathbf{S}_{\mu} \right)$$
$$\begin{bmatrix} \log(\sigma_{i,1,1}) \\ \log(\sigma_{i,1,2}) \\ \log(\sigma_{i,2,1}) \\ \log(\sigma_{i,2,2}) \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} \sigma_{\text{mean},1,1} \\ \sigma_{\text{mean},1,2} \\ \sigma_{\text{mean},2,1} \\ \sigma_{\text{mean},2,2} \end{bmatrix}, \mathbf{S}_{\sigma} \right) \quad \text{(S20)}$$

With this specification the covariance and correlation matrices are now 4x4 as opposed to 2x2. We used the same prior distributions as in the other two group-level model specifications (i.e. the specification in the main text and the Joint Separate specification described above). We term this model the "Joint Single" model, given that it assumes that all person-level parameters of the same type arise from a single group-level multivariate normal distribution. Unlike the other models, Equation S20 does not directly estimate the test-retest correlation of the difference in parameters between conditions. Therefore, we computed the test-retest correlation post-hoc using the MCMC samples. Specifically, for each sample s, we: (1) computed the difference in person-level parameters between conditions for each person and session (e.g., $\mu_{i,\Delta,1} = \mu_{i,2,1} - \mu_{i,1,1}$; $\sigma_{i,\Delta,1} = \log(\sigma_{i,2,1}) - \log(\sigma_{i,1,1})$), and then (2) computed the Pearson's correlation between the differences from (1) for each parameter across participants between the two sessions (e.g., $\text{cor}(\mu_{1:N,\Delta,1},\ \mu_{1:N,\Delta,2})$; $\text{cor}(\sigma_{1:N,\Delta,1},\ \sigma_{1:N,\Delta,2})$). This procedure results in a posterior distribution of test-retest correlations for each of the parameter differences (i.e. the "Stroop effects" in the context of the Stroop task), which we interpreted in the same way as those directly estimated in

the other two models. Figure S9 shows the test-retest correlations estimated using the Joint Single models. Like the Joint Separate models, the results were not notably different from the model presented in the main text (c.f. Figure S9 to Figure 6).

Figure S9. Test-retest correlations for all tasks using the Joint Single models.

Here we show means and 95% confidence intervals for the two-stage summary approach (for both sample mean and standard deviation contrasts) in black, along with the posterior means and 95% highest density intervals for the generative model parameter estimates (in various shades of red). The Implicit Association Test (IAT) datasets are from the Self-Concept (introversion/extraversion; Study 1a) and Race (Black/White; Study 2b) versions. This model uses the group-level specification described in Equation S20.



8. A Generative Solution for Outlier Response Times

The Normal, Lognormal, and Shifted Lognormal models used throughout the main text are unable to capture "outlier" response times that may arise naturally when people initiate a response either before processing a stimulus or after a long period of losing focus on the task. Typically, such responses are removed from the data before fitting a model using cutoffs that are grounded in domain knowledge, yet still chosen in a heuristic way. For example, 100ms is often used as a conservative lower bound on allowable response times—times faster than 100ms are almost certainly not reflecting the cognitive process of interest, but we will likely retain some "contamination" responses that occur past 100ms. As noted in the main text, a researchers' choice of lower and upper bounds for filtering out contamination response times can have a large effect on the resulting inference (Parsons, 2020). Given our decision to retain all response times greater than 0ms, it is natural to ask whether this choice can unduly affect reliability results, and how one might build a generative model to account for contamination response times without the need for arbitrary cutoffs.

To begin, we can assume that each response is generated from either the cognitive process of interest or from a contamination process. Statistically, this type of specification is known as a mixture model. To build on our earlier models, we will assume that the cognitive process is represented by the Shifted Lognormal distribution. For the contamination process, we need to choose a distribution that ranges from the lowest to the highest response time that we reasonably expect to observe in the experimental data. In our case, we chose a uniform distribution ranging from 0 ms to the largest response time that occurs in the experimental data across participants. The uniform distribution is a suitable choice here because it can capture response times that are

either faster or slower than what we would expect to arise naturally.

Mathematically, the contamination mixture model is defined as

$$p(\mathrm{RT}_{i,c,t}) = \begin{cases} \mathcal{U}(0, \max(\mathbf{RT}), \text{ if } \mathrm{RT}_{i,c,t} < \delta_{i,t} \\ \mathcal{U}(0, \max(\mathbf{RT}) \cdot \lambda_i + \mathcal{SLN}(\delta_{i,t}, \mu_{i,c,t}, \sigma_{i,c,t}) \cdot (1 - \lambda_i), \text{ otherwise} \end{cases} \tag{S21}$$

where $\mathcal{SLN}$ is the Shifted Lognormal from equation 4 in the main text (with the same parameter interpretations), $\mathcal{U}$ is the uniform distribution, and λ_i is the contamination mixture parameter, which indicates the probability that a response is generated per the uniform contamination distribution as opposed to the Shifted Lognormal distribution. We imposed a strong prior on λ_i that implies only ~5% of observed responses should arise from the contamination process:

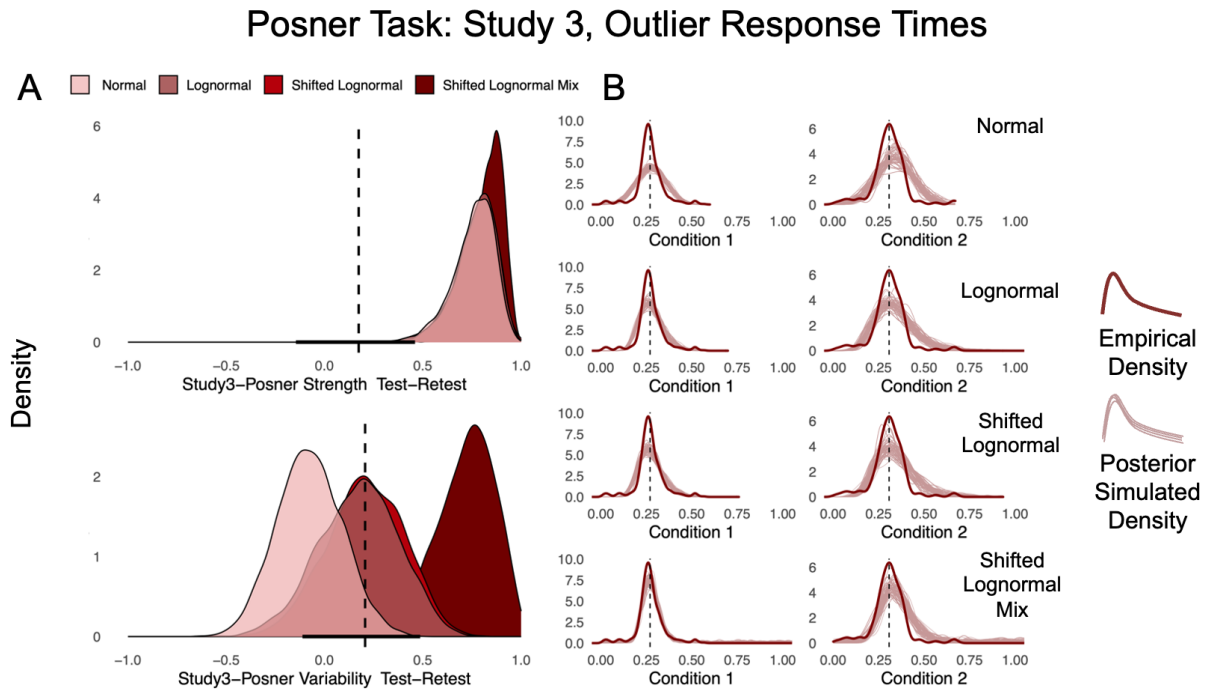
$$\begin{aligned} \Phi^{-1}(\boldsymbol{\lambda}) &\sim \mathcal{N}(\lambda_{\mathrm{mean}}, \lambda_{\mathrm{sd}}) \\ \lambda_{\mathrm{mean}} &\sim \mathcal{N}(-1.65, 0.1) \\ \lambda_{\mathrm{sd}} &\sim \text{half-}\mathcal{N}(0, 0.2) \end{aligned} \tag{S22}$$

Above, Φ^{-1} is the inverse of the standard normal cumulative distribution function, which ensures that the mixing probability λ_i is between 0 and 1.

To illustrate the power of this relatively simple extension of the Shifted Lognormal model, we refit the model to the data from the Posner task described in section 6.3 above. In this analysis, we also refit the Normal, Lognormal, and Shifted Lognormal to the data after first filtering out all response times less than 100 milliseconds to facilitate a direct evaluation of the effectiveness of the contamination parameter λ. We chose the Posner task because it contains many contamination responses that cause poor fit for the Shifted Lognormal model (see Figure S5). Without the contamination process, the shift (δ) parameter in the model must be at least as low as the fastest response time, so even a single ~50ms response time causes the model to generate unreasonably low response time predictions, which inflates the variance parameter in the model. Figure S10 shows that the contamination mixture model described in equations S21-

22 alleviates this misfit issue a great deal, even when compared to the other three models using a 100 millisecond lower bound cutoff to filter out rapid response times. In comparing Figure S10 to Figure S5, we can see that although the 100-millisecond filtering of fast response times improved model fits, the more substantial improvement was seen with the inclusion of the contaminant parameter λ. In addition to better predictive performance, the contamination mixture model also leads to a much higher estimate of the test-retest correlation of the sigma parameter in the model. Altogether, these results further bolster our primary message—that appropriately selected models of the generative process allow us to learn from our data in ways that heuristic procedures do not.

Figure S10. Test-retest correlations and model misfit for the Posner Cueing task. (A) Posterior distributions for test-retest correlations of each of the four generative models (including the contamination mixture model). (B) Posterior predictive simulations and sample means for each of the generative models for a representative participant. Models from top to bottom are the Normal, Lognormal, Shifted Lognormal, and Shifted Lognormal contamination mixture models. Note that in this analysis, all models except for the contaminant mixture model have filtered out all response times less than 100 milliseconds before fitting (cf. Figure S5 where no cutoff was used).



Supplementary References

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Teater, J. E., Myung, J. I., & Pitt, M. A. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports.* Manuscript accepted for publication.

Betancourt, M., & Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. *arXiv* 1312.0906. http://arxiv.org/abs/1312.0906

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124, 1-20. doi:10.1037/rev0000047

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin and Review, 18*, 204-210. doi:10.3758/s13423-010-0030-4

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990). On the control of automatic processes: A parallel-distributed processing account of the Stroop effect. *Psychological Review, 97*, 332-361.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469-487. doi:10.1037/0022-3514.89.4.469

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology, 75*, 485-491. doi:10.2307/1419876

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit andexplicit measures. *Personality and Social Psychology Bulletin* 43: 300-312.doi:10.1177/0146167216684131

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7: 457-472. doi:10.2307/2246093

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *PsyArXiv preprint*, 1-13. doi:10.31234/osf.io/rybh9

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions:an example using the Stroop task. *Psychological Bulletin, 109*, 340-347. doi:10.1037/0033-2909.109.2.340

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* 103: 1-21. doi:10.3758/s13428-017-0935-1

Hockley, W. E., & Corballis, M. C. (1982). Tests of serial scanning in item recognition. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 36*, 189-212. doi:10.1037/h0080637

Jepma, M., Wagenmakers, E. J., & Nieuwenhuis, S. (2012). Temporal expectation and information processing: A model-based analysis. *Cognition, 122*, 426-441. doi:10.1016/j.cognition.2011.11.014

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Social Psychological and Personality Science, 8*, 413-423. doi:10.1177/1948550617703174

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards?, *157*, 126-138. doi:10.1016/j.cognition.2016.08.020

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93*, 353-368. doi:10.1037/0022-3514.93.3.353

Kvam, P. D. (2019). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology, 91*, 14-37. doi:10.1016/j.jmp.2019.03.001

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review.* Advance online publication. doi:10.1037/rev0000215

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica, 104*, 167-190. doi:10.1016/s0001-6918(00)00019-6

Luckman, A., Donkin, C., & Ben R Newell. (2017). Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice. *Psychonomic Bulletin and Review, 25*, 785-792. doi:10.3758/s13423-017-1330-8

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology, 57*, 53-67. doi:10.1016/j.jmp.2013.05.005

Odum, A. L. (2011). Delay discounting: I'm a k, you're a k. *Journal of the Experimental Analysis of Behavior* 96: 427-439. doi:10.1901/jeab.2011.96423

Parsons, S. (2020).Exploring reliability heterogeneity with multiverse analyses: Dataprocessing decisions unpredictably influence measurement reliability. PsyArXiv preprint. doi:10.31234/osf.io/y6tcz

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika, 80*, 491-513. doi:10.1007/s11336-013-9396-3

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* 15: 72-101. doi:10.2307/1412159

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review, 125*, 329-362. doi:10.1037/rev0000089

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*, 385-402. doi:10.1027/1618-3169/a000218

Wagenmakers, E. J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model forresponse time and accuracy. *Psychonomic Bulletin and Review, 14*, 3-22. doi:10.3758/BF03194023

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record, 58*, 475-482. doi:10.1007/BF03395630

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the Flanker Task: Discrete versus gradual attentional selection. *Cognitive Psychology, 63*, 210-238. doi:10.1016/j.cogpsych.2011.08.001

Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). ADOpy: A Python package for adaptive design optimization. *In press at Behavior Research Methods*. doi:10.31234/osf.io/mdu23